On the Morse-Hedlund complexity gap

Julien Cassaigne   François Nicolas

December 4, 2009

Abstract

In 1938, Morse and Hedlund proved that the subword complexity function of a two-sided infinite word is either bounded or at least linearly growing. In 1982, Ehrenfeucht and Rozenberg proved that this gap property holds for the subword complexity function of any language. Their result was then sharpened in 2005 by Balogh and Bollobás. The aim of the present paper is to present a self-contained, compact proof of Ehrenfeucht and Rozenberg's result.

1 Notation and definitions

The semiring of natural integers is denoted $\mathbb{N}$. The ring of rational integers is denoted $\mathbb{Z}$.

Words and languages. Throughout this paper A denotes a finite set of symbols and is called the *alphabet*. A *word* over A is a finite string of elements of A. The set of all words over A is denoted $A^\star$. A *language* over A is a subset of $A^\star$. For every $w \in A^\star$, the *length* of w is denoted $|w|$. For each $n \in \mathbb{N}$, the set of all n-length words over A is denoted A^n. The word of length zero is called the *empty word*. The set of all non-empty words over A is denoted A^+. Word concatenation is denoted multiplicatively.

Factor and complexity function. Let $x \in A^\star$, $y \in A^\star$, and $L \subseteq A^\star$. We say that x is a *prefix* (resp. a *suffix*) of y if there exists $w \in A^\star$ such that $y = xw$ (resp. $y = wx$). We say that x is a *factor* of y if there exist w, $w' \in A^\star$ such that $y = wxw'$. By extension, we say that x is a factor of L if x is a factor of some word in L. For each $n \in \mathbb{N}$, $F_n(L)$ denotes the set of all n-length factors of L. The function from $\mathbb{N}$ to itself that maps each $n \in \mathbb{N}$ to the cardinality of $F_n(L)$ is called the *complexity function* of L. Note that the complexity function of any non-empty language maps 0 to 1 because the empty word is a factor of every word. Note also that a language has the same complexity function as its set of factors.

Infinite and bi-infinite words. A *(right-)infinite word* over A is a function from $\mathbb{N}$ to A. A *bi-infinite* word over A is a function from $\mathbb{Z}$ to A. Let u be an infinite or bi-infinite word over A. We say that x is a factor of u if there exists i in the domain of u such that $x = u(i)u(i+1)u(i+2)\cdots u(i+|x|-1)$. The set of all factors of u is called the language of u. The complexity function of u is defined as the complexity function of its language.

2 The Morse-Hedlund complexity gap

The aim of this paper is to present a self-contained, compact proof of:

Theorem 1 (Ehrenfeucht and Rozenberg, 1982 [4]). *Let p be the complexity function of some language. Either $p(n)$ is greater than n for every $n \in \mathbb{N}$, or p is bounded from above.*

For instance, it follows from Theorem 1 that no complexity function grows like $\sqrt{n}$. However, the lower bound is trivially achievable.

Example 1. Consider the language $U = \left\{\mathtt{a}^i\mathtt{b}^j : i, j \in \mathbb{N}\right\}$. For each $n \in \mathbb{N}$, it is clear that $F_n(U) = \left\{\mathtt{a}^{n-k}\mathtt{b}^k : k = 0, 1, 2, \ldots, n\right\}$, so the complexity function of U maps n to $n+1$.

In addition to proving Theorem 1, Ehrenfeucht and Rozenberg described the class of those languages with bounded complexity functions:

Theorem 2 (Ehrenfeucht and Rozenberg, 1982 [4]). *Let L be a language. The complexity function of L is bounded from above if, and only if, there exists a finite subset $T \subseteq A^\star \times A^\star \times A^\star$ such that*

$$L \subseteq \bigcup_{(x,y,z)\in T} \{xy^n z : n \in \mathbb{N}\} \ .$$

Before proving Theorem 1 in Section 3, let us state the various related results that can be found in the literature. Everything starts with Morse and Hedlund's celebrated papers [6] and [7]. They were published in 1938 and 1940, respectively.

Definition 1. *We say that a function $p\colon \mathbb{N} \to \mathbb{N}$ is* FIATC *(first increasing and then constant) if there exists $m \in \mathbb{N}$ such that*

- *$p(0) < p(1) < p(2) < \cdots < p(m)$ and*
- *$p(m+n) = p(m)$ for every $n \in \mathbb{N}$.*

Theorem 3 (More and Hedlund, 1938 [6, 3]). *Let u be a bi-infinite word and let p denote the complexity function of u.*

- *If u is not periodic then p is increasing.*
- *If u is periodic then p is* FIATC *and $\sup_{n\in\mathbb{N}} p(n)$ is the least period of u.*

On the one hand, FIATC functions are clearly bounded. On the other hand, observe that any increasing function $p\colon \mathbb{N} \to \mathbb{N}$ satisfies $p(n) \geq n + p(0)$ for every $n \in \mathbb{N}$. Hence, Theorem 3 implies that Theorem 1 holds for the particular case where p is the complexity function of a bi-infinite word. Again, the lower bound is achievable:

Example 2. Consider the bi-infinite word u over $\{\mathtt{a}, \mathtt{b}\}$ given by: $u(i) = \mathtt{b}$ for every $i \in \mathbb{N}$ and $u(-i) = \mathtt{a}$ for every $i \in \mathbb{N} \setminus \{0\}$. The language of u equals U, where U is as in Example 1. Therefore, the complexity function of u maps n to $n + 1$ for each $n \in \mathbb{N}$.

As illustrated with the following two examples, bounded complexity functions are not necessarily FIATC and unbounded complexity functions are not necessarily increasing:

Example 3. Let p denote the complexity function of $L = \left\{\mathtt{ba}^{2k}\mathtt{b} : k \in \mathbb{N}\right\}$. For each integer $n \geq 1$, let $X_n = \{\mathtt{a}^n, \mathtt{ba}^{n-1}, \mathtt{a}^{n-1}\mathtt{b}\}$. If n is an odd positive integer then $F_n(L) = X_n$ and thus $p(n) = 3$. If n is an even positive integer then $F_n(L) = X_n \cup \{\mathtt{ba}^{n-2}\mathtt{b}\}$ and thus $p(n) = 4$.

Example 4. Let U be as in Example 1. Let p denote the complexity function of $L = U \cup \left\{\mathtt{ab}^{2k}\mathtt{a} : k \in \mathbb{N}\right\} \cup \left\{\mathtt{ba}^{2k}\mathtt{b} : k \in \mathbb{N}\right\}$. If $n = 2$ or if n is an odd positive integer then $F_n(L) = F_n(U)$ and thus $p(n) = n + 1$. If n is an even integer greater than or equal to 4 then $F_n(L) = F_n(U) \cup \{\mathtt{ab}^{n-2}\mathtt{a}, \mathtt{ba}^{n-2}\mathtt{b}\}$ and thus $p(n) = n + 3$.

The most famous variant of Theorems 1/2 and 3 is:

Theorem 4 ([3], see also [2, 5, 8])**.** *Let u be an infinite word and let p denote the complexity function of u.*

- *If u is not eventually periodic then p is increasing.*
- *If u is eventually periodic then p is* FIATC *and the period of u is not greater than $\sup_{n \in \mathbb{N}} p(n)$.*

An infinite word is called *Sturmian* if its complexity function maps each $n \in \mathbb{N}$ to $n + 1$: by Theorem 4, Sturmian words are those non-eventually-periodic infinite words with minimum complexity. There is no trivial example of Sturmian word. The study of Sturmian words was initiated by Morse and Hedlund in 1940 [7]. It is still an active field of research [5, 8].

The next result is the latest improvement of Theorem 1.

Theorem 5 (Balogh and Bollobás, 2005 [1])**.** *Let ϕ be the function that maps each real number x to* $\left\lceil \frac{x+1}{2} \right\rceil \left\lfloor \frac{x+1}{2} \right\rfloor$.

- *Let p be the complexity function of some language and let $m \in \mathbb{N}$. If $p(m) \leq m$ then $p(n + p(m) + m) \leq \phi(p(m))$ for every $n \in \mathbb{N}$.*
- *For each $k \in \mathbb{N}$, there exists a function $p_k\colon \mathbb{N} \to \mathbb{N}$ such that p_k is the complexity function of some binary language and both sets $\{n \in \mathbb{N} : p_k(n) = k\}$ and $\{n \in \mathbb{N} : p_k(n) = \phi(k)\}$ are infinite.*

The second part of Theorem 5 ensures that the function ϕ is optimal.

3 The proof of Theorem 1

In what follows, L denotes a language over the alphabet A and p denotes the complexity function of L.

Definition 2 (Special factor)**.** *We say that a word $w \in A^\star$ is a* special factor *of the language L if there exist a, $b \in A$ with $a \neq b$ such that both wa and wb are factors of L.*

Let $\rho\colon A^+ \to A^\star$ be the function mapping each non-empty word $w \in A^+$ to its $(|w|-1)$-length prefix: for each $w \in A^+$, there exists $a \in A$ such that $w = \rho(w)a$. For any language $L \subseteq A^\star$ and any $n \in \mathbb{N}$, ρ maps each word in $F_{n+1}(L)$ to a word in $F_n(L)$. A word $w \in A^\star$ is a special factor of L if, and only, if ρ maps to w more than one factor of L. It follows that L admits an n-length special factor if, and only if, ρ is not injective on $F_{n+1}(L)$.

Lemma 1. *If a language only admits finitely many special factors then its complexity function is eventually constant.*

Proof. Assume that L only admits finitely many special factors. Let $n \in \mathbb{N}$ be such that L does not admit any n-length special factor. Then, ρ induces an injection from $F_{n+1}(L)$ into $F_n(L)$. Inequality $p(n+1) \leq p(n)$ follows. Therefore, p is non-increasing on $\{n \in \mathbb{N} : n > m\}$, where $m \in \mathbb{N}$ is denotes the maximum length of a special factor of L. $\square$

Note that the converse of Lemma 1 does not hold in general:

Example 5. Consider the case where $L = \{\mathtt{a}^k\mathtt{b} : k \in \mathbb{N}\}$. The complexity function of L is eventually constant: for every integer $n \geq 1$, $F_n(L) = \{\mathtt{a}^n, \mathtt{a}^{n-1}\mathtt{b}\}$ and thus $p(n) = 2$. However, L admits infinitely many special factors: for every $n \in \mathbb{N}$, $\mathtt{a}^n$ is a special factor of L.

Exercise 1. Prove that if the language of an infinite word admits only finitely many special factors then this infinite word is eventually periodic.

Exercise 2. Prove that if the language L only admits finitely many special factors then there exists a finite subset $T \subseteq A^\star \times A^\star$ such that

$$L \subseteq \bigcup_{(x,y) \in T} \{xy^n : n \in \mathbb{N}\} \ .$$

Definition 3. *We say that the language L is* (right-)extendable *if for each $w \in L$, there exists $a \in A$ such that $wa \in L$.*

Example 6. The language of any infinite or bi-infinite word is extendable.

Lemma 2. *If a language admits infinitely many special factors then its complexity function is increasing.*

Proof. Let $n \in \mathbb{N}$. If L is extendable then ρ induces a surjection from $F_{n+1}(L)$ onto $F_n(L)$. If L admits infinitely many special factors then L admits a special factor of length n because every suffix of a special factor is also a special factor, and thus ρ is not injective on $F_{n+1}(L)$. Hence, if L is extendable and admits infinitely many special factors then ρ induces a non-bijective surjection from $F_{n+1}(L)$ onto $F_n(L)$, and thus inequality $p(n+1) > p(n)$ holds. $\square$

Exercise 3. Prove that the complexity function of any extendable language is either increasing or FIATC.

Exercise 4. For each $n \in \mathbb{N}$, let $s(n)$ denote the number of n-length special factors of L.

1. Let α denote the cardinality of A. Prove that $p(n+1) - p(n) \leq (\alpha - 1)s(n)$ for every $n \in \mathbb{N}$.

2. Prove that if L is extendable then $p(n+1) - p(n) \geq s(n)$ for every $n \in \mathbb{N}$.

Lemmas 1 and 2 can be easily deduced from questions 1 and 2 of Exercise 4, respectively.

Lemma 3. *Let X and Y be two languages such that Y is finite and non-empty. The complexity function of XY is bounded if, and only if, the complexity function of X is bounded.*

Proof. Let f, g and h denote complexity functions of X, Y, and XY, respectively. We have to prove that h is bounded if, and only if, f is bounded.

Since Y is non-empty, every factor of X is also a factor of XY, and thus f is bounded from above by h. The "only if part" follows.

Let us now prove the "if part". The inclusion

$$F_n(XY) \subseteq \bigcup_{k=0}^{n} F_{n-k}(X)F_k(Y)\,,$$

yields the inequality

$$h(n) \leq \sum_{k=0}^{n} f(n-k)g(k)\,. \tag{1}$$

Let $M = \sup_{n \in \mathbb{N}} f(n)$ and $S = \sum_{n \in \mathbb{N}} g(n)$. Observe that $S < \infty$. Assume that f is bounded. Then we have $M < \infty$ and it follows from Equation (1) that

$$h(n) \leq M \sum_{k=0}^{n} g(k) \leq MS\,.$$

Hence, MS is a finite upper bound for h. $\square$

We can now prove Theorem 1.

Proof of Theorem 1. For each $k \in \mathbb{N}$, let L_k denote the set of all $w \in A^\star$ such that $wA^k \cap L \neq \emptyset$. Let L' denote the set of all $w \in A^\star$ such that w is a factor of L_k for every $k \in \mathbb{N}$. Let p' denote the complexity function of L', and for each $k \in \mathbb{N}$, let p_k denote the complexity function of L_k. Since each word in L' is a factor of $L_0 = L$, p' bounds p from below. Therefore, if p' is increasing then for every $n \in \mathbb{N}$, it holds that $p(n) \geq p'(n) > n$. It remains to show that if p' is not increasing then p is bounded.

Claim 1. *For any i, $j \in \mathbb{N}$ with $i \leq j$, all factors of L_j are factors of L_i.*

Proof. Each word in L_j is a prefix of a word in L_i: for each $w \in L_j$, there exists $x \in A^{j-i}$ such that $wx \in L_i$. □

Claim 2. *The language L' is extendable.*

Proof. For each $w \in L_{k+1}$, there exists $a \in A$ such that $wa \in L_k$. Therefore, for each factor w of L_{k+1}, there exists $a \in A$ such that wa is a factor of L_k. Let $w \in L'$. For each $k \in \mathbb{N}$, let $a_k \in A$ be such that wa_k is a factor of L_k. The finite alphabet A contains a letter a such that $a_k = a$ for infinitely many $k \in \mathbb{N}$. Therefore, wa is a factor of L_k for infinitely many $k \in \mathbb{N}$. It now follows from Claim 1 that $wa \in L'$. □

Claim 3. *If p_k is bounded for some $k \in \mathbb{N}$ then p is bounded.*

Proof. Remark that L is a subset of $L_k A^k \cup \{w \in L : |w| < k\}$. Since $\{w \in L : |w| < k\}$ is finite, p is bounded whenever the complexity function of $L_k A^k$ is bounded. Besides, Lemma 3 (applied with $X = L_k$ and $Y = A^k$) ensures that the complexity function of $L_k A^k$ is bounded whenever p_k is bounded. □

Claim 4. *For each $n \in \mathbb{N}$, there exists $k \in \mathbb{N}$ such that $F_n(L_k) \subseteq L'$.*

Proof. Claim 1 ensures

$$F_n(L_0) \supseteq F_n(L_1) \supseteq F_n(L_2) \supseteq F_n(L_3) \supseteq \cdots$$

and since all sets are finite there exists $k \in \mathbb{N}$ such that

$$F_n(L_k) = F_n(L_{k+1}) = F_n(L_{k+2}) = F_n(L_{k+3}) = \cdots$$

□

Assume that p' is not increasing. Then, combining Claim 2 and Lemma 2, we get that L' only admits finitely many special factors. Let $n \in \mathbb{N}$ be greater than the length of every special factor of L'. By Claim 4, there exists $k \in \mathbb{N}$ such that $F_n(L_k) \subseteq L'$. Clearly, L_k has no special factor with length $n-1$ or more, and thus p_k is eventually constant by Lemma 1. It now follows from Claim 3 that p is bounded. □

Exercise 5. Prove Theorem 2 (Hint: use Exercise 2).

References

[1] J. Balogh and B. Bollobás. Hereditary properties of words. *Informatique Théorique et Applications*, 39(1):49–65, 2005.

[2] L. E. Bush. The William Lowell Putnam mathematical competition. *The American Mathematical Monthly*, 62(8):558–564, 1955.

[3] E. M. Coven and G. A. Hedlund. Sequences with minimal block growth. *Mathematical Systems Theory*, 7(2):138–153, 1973.

[4] A. Ehrenfeucht and G. Rozenberg. On subword complexities of homomorphic images of languages. *Informatique Théorique et Applications*, 16(4):303–316, 1982.

[5] M. Lothaire. *Algebraic combinatorics on words*. Number 90 in Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2002.

[6] M. Morse and G. A. Hedlund. Symbolic dynamics. *American Journal of Mathematics*, 60(4):815–866, 1938.

[7] M. Morse and G. A. Hedlund. Symbolic dynamics II. Sturmian trajectories. *American Journal of Mathematics*, 62(1):1–42, 1940.

[8] N. Pytheas Fogg. *Substitutions in Dynamics, Arithmetics and Combinatorics*, volume 1794 of *Lecture Notes in Mathematics*. Springer, 2002. Edited by V. Berthé, S. Ferenczi, C. Mauduit and A. Siegel.